Fostering Serendipity through Big Linked Data

Muhammad Saleem[1], Maulik R. Kamdar[2], Aftab Iqbal[2], Shanmukha Sampath[2], Helena F. Deus[3], and Axel-Cyrille Ngonga[1]

[1] Universität Leipzig, IFI/AKSW, PO 100920, D-04009 Leipzig
lastname@informatik.uni-leipzig.de
[2] Digital Enterprise Research Institute, National University of Ireland, Galway.
{firstname.lastname}@deri.org
[3] Foundation Medicine Inc. One Kendal Square Cambridge, MA
hdeus@foundationmedicine.com

Abstract. The amount of bio-medical data available over the Web grows exponentially with time. The large volume of the currently available data makes it difficult to explore, while the velocity at which this data changes and the variety of formats in which bio-medical is published makes it difficult to access them in an integrated form. Moreover, the lack of an integrated vocabulary makes querying this data difficult. In this paper, we advocate the use of Linked Data to integrate, query and visualize big bio-medical data. As a proof of concept, we show how the constant flow of bio-medical publications can be integrated with the 7.36 billion large Linked Cancer Genome Atlas dataset (TCGA). Then, we show how we can harness the value hidden in that data by making it easy to explore within a browsing interface. We evaluate the scalability of our approach by comparing the query execution time of our system with that of FedX on Linked TCGA.

Keywords: TCGA, PubMed, RDF

1 Introduction

Over the last years, the amount of Linked Data published has grown significantly. Especially the bio-medical data available as RDF is comprised in partly very large datasets, one of the newest additions to this family of datasets being Linked TCGA [2], a 7.6-billion-triples-strong dataset. Making bio-medical data available as Linked Data presents the obvious advantage of easing the integration of this data, which promises to support bio-medical experts during the analysis of, exploration of and extraction of novel knowledge from this data. Yet, the necessary data management solutions for RDF data still need to be perfected to obtain scalable integrated solutions that can deal with Big Linked Data, i.e., Linked Data which display the three main characteristics of Big Data (volume, velocity and value).

In this paper, we present a scalable approach that aims to support the serendipitous discovery of bio-medical hypotheses by providing an interface for

the analysis and exploration of Big Linked Data. The back-end of our application supports the management and querying of high volumes of Linked Data as well as the continuous integration of this data with novel data from external data streams. Here, we consider especially Linked TCGA and its continuous integration with RDF data extracted from the semi-structured and unstructured content of PubMed. We chose PubMed because it contains more than 23 million publications and provide an interface that allows discovering novel publications as soon as they are made available. The user interface developed on top of the resulting dataset presents an easily understandable, integrated, up-to-date view of the information available in the back-end. The intuition behind our work is that when presented with such an interface, experts can detect unexpected correlations amongst known resources. These unexpected correlations can then form the basis for a serendipitous discovery. For example, bio-medical experts could detect that cancers of type A tend to metastasize into cancers of a type B, leading to the question of why this particular cell migration occurs. This question could then lead to the serendipitous formulation of new research questions, e.g., pertaining to the rheology of certain cancer types.

In the following, we present the datasets underlying our implementation. We then give an overview of the architecture of our tool and show how it support volume, velocity and variability to generate novel value from large datasets. We then present the user interface of our tool as well as possible usages of our interface. The evaluation section shows that our data infrastructure outperforms the state of the art in the management of large amounts of data while the conclusion gives insights in future work.

2 Linked TCGA

The aim of TCGA project is to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. Furthermore, the TCGA data portal provides public access[4] to cancer data in order to enable researchers perform cancer analysis on real data. Currently, the said portal contains data for 30 different cancer types collected from 9000 patients[5]. To data, 21 data files are collected for each patient, leading to a total of 147,645 raw data files (12.7 TB), of which 53,694 contain processed data [2]. According to information in the TCGA portal, this is only 46% of the expected data with new data being submitted every day.[6]

Linked TCGA is the RDFized version of the Cancer Genome Atlas presented in [2]. The main aim of this work is to facilitate the querying and live integration of TCGA from multiple sources via remote SPARQL query processing. The total estimated size of the Linked TCGA data is over 30 billion triples [2]. In this work, we use a total of 7.36 billions triples of Linked TCGA obtained from 10 tumours. The details about data is given is in Table 1.

[4] `https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp`
[5] `https://tcga-data.nci.nih.gov/tcga/`
[6] `http://cancergenome.nih.gov/`

Tumor Type	Original Size(GB)	Refined Size (GB)	RDFized Size (GB)	Triples (Million)
Cervical (CESC)	8.75	2.44	8.86	400.19
Rectal adenocarcinoma (READ)	8.07	2.25	9.04	413.31
Papillary Kidney (KIRP)	10.40	2.90	10.4	469.65
Bladder cancer (BLCA)	12.16	3.39	12.3	556.38
Acute Myeloid Leukemia (LAML)	14.85	4.14	15.1	684.05
Lower Grade Glioma (LGG)	17.08	4.76	17.1	778.82
Prostate adenocarcinoma (PRAD)	18.05	5.03	18.1	821.01
Lung squamous carcinoma (LUSC)	20.63	5.75	20.5	927.08
Cutaneous melanoma (SKCM)	23.22	6.47	23.2	1050.94
Head and neck squamous cell(HNSC)	27.6	7.69	27.5	1245.37

Table 1: Overview of the TCGA data used by our approach.

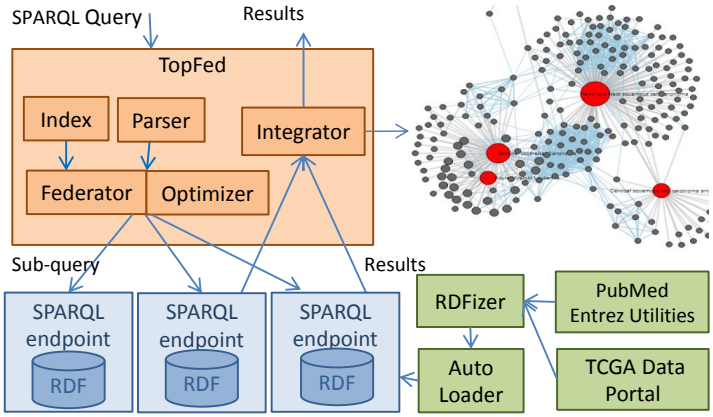

Fig. 1: General architecture of the proposed system

3 General architecture of our solution

The general architecture of our system is shown in Figure 1 and is explained in the following sections.

3.1 TopFed

The backbone of our approach is the TopFed[7], a Linked TCGA federated SPARQL query processing engine specially designed for efficient integration of data from multiple TCGA SPARQL endpoints. The insight behind this engine is to use the intelligent distribution of data to reduce the number of sources selected (without losing the recall) for processing federated SPARQL queries. By selecting fewer sources than state-of-the-art approaches such as FedX [3], our approach can compute the answer to queries significantly faster, leading to acceptable response time for large Linked TCGA SPARQL queries.

3.2 Integration of publication data

We have considered a list of keywords related to TCGA tumors and search PubMed articles using the Entrez Programming Utilities [8] which return a list of

[7] `https://code.google.com/p/topfed/`
[8] `http://www.ncbi.nlm.nih.gov/books/NBK25501/`

PubMed publication IDs. Once we got the list of article IDs which are tagged with the keywords, we used the E-utilities again to retrieve meta information about each article by providing the article ID as input. The response of the query, returned in XML format, is then converted to RDF using our custom written script and continuously loaded into the TCGA SPARQL endpoints.

3.3 Visualization

One of the most prominent challenges in delivering and using data-driven solutions for any type of human process is the provision of data visualization/-summarization tools that are intuitive and easy to use for the experts at whom the datasets are targeted. Summarizing and displaying the evidence that these experts need for making informed decisions, reusing the results in new contexts and addressing their own challenges is the main task of such data visualization tools. The current method by which physicians look for information on the web is through peer-reviewed publications. However, with the indexing of over 10000 papers in PubMed every year, keeping current with the literature and using this knowledge to derive new research questions has become a herculean task.

To facilitate the intuitive exploration of the information available in TCGA datasets and tumor-related publications, we devised an interactive Visual Analytics Platform (Figure 2). The platform is composed of two panels. The main panel features a highly dense, force-directed network graph linking the different tumor typologies analysed in TCGA to the publication resources where more information about these tumors can be discovered. The linking/display method associates two publications if they have more than 5 common Mesh Terms and are linked using blue-colored edges. Selecting a publication node presents the metadata of the publication (author, abstract, mesh terms, chemicals cited, etc.) and a link to the original PubMed page. On the other hand, selecting a tumor typology node will aggregate the data collected on different cancer patients with that tumour type.

In addition to the tumour data, we also display methylation data as this phenomena is known to be highly relevant for cancer progression detection. Methylation patterns in cancer cells are known to reflect the silencing or "turning-off" of cancer protecting genes (i.e. tumour suppressor genes), thus allowing the cancer to progress. These results are visualized as scatter plots in a tabbed interface and reflect the amount of methylation in each genomic position, where the chromosomes are indicated against the Y-axis and the specific position in the chromosome is on the X-axis. The circles reflect the positions in the genome where methylation was uniquely detected in the cancer cells, and the size of the circles is directly proportional to the amount (beta value) of methylation in that region. The interface allows the simultaneous comparison of these results between different patients and also links the underlying methylated gene to the publication(s) which mentions it. A prototype of the interface is available at `http://srvgal78.deri.ie/tcga-pubmed/`.

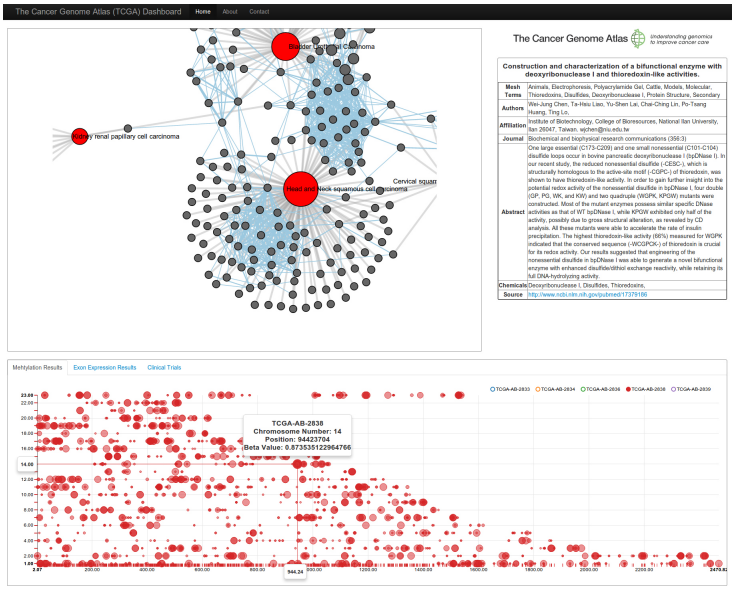

Fig. 2: Visual Analytics Platform for the Integrated Visualization of TCGA methylation datasets and related PubMed resources

4 Use Cases

Our framework enables a variety of use cases, of which two are explained below.

4.1 Enabling Evidence-based genomic medicine

Data from TCGA is of high value for oncologists as it enables matching the evidence that they find for their own patients with those enrolled in the TCGA project, including both clinical and genomic sets. It is well known that specific genomic alterations in each individual's cancer affect response to treatment and sensitivity to drugs. As such, a physician could, for example, use our visualization to compare their own patient's methylation patterns against that for other patients enrolled in TCGA. Since genomic information in TCGA is linked with each patient's clinical prognostic and follow-up, the physician could assert, based on the similarity of genomic results, whether a patient would respond well to a given drug by observing the other patient's reaction. What this also enables is medical decisions that are highly informed by the evidence. Cancer, we now know, is a genetic disease. This means that the location where the tumour occurs (e.g. brain, liver, etc) is less relevant for its treatment than the genetic signature that the cancer cells express (i.e. whether genes are silenced, amplified, etc). However, drugs that are approved by regulatory agencies, and many publication resources, are still approved in the context of a single tumour typology. By making use of cross-resource linking, we enable the discovery on whether a drug could be applied to more than on tumour typology, by linking the two typologies

through their genomic signature. As an example, a publication resource that is linked to two or more tumor typologies may mean that a discovery has been made that affects both cancer typologies and therefore the same drug or set of drugs may be applied.

4.2 Generation of new hypotheses

In addition to aiding evidence based genomic medicine, the availability of this type of linked information can also facilitate inter-disciplinary research. Some types of cancer (e.g. breast cancer) are more common than others and therefore the intricacies of their genomic signatures and genetic events tend to be more well known. However, for many rare cancers (e.g. pancreatic cancer), knowledge is more scattered and harder to find. The resource that we make available will enable researchers in the less common tumor typologies to discover association between their cancer of interest, and those that are more well studied. By finding papers where two tumor typologies co-occur, a researcher can hypothesize that the treatments and genomic events that are valid and have been proved to be relevant in the most common type of cancer, may also be relevant in the less common tumor typology. They can then exploit the genomic data in both cases to support or reject this hypothesis.

Another possible arena for hypothesis generation is that of tumour cell migration. Cancer experts have shown in the past that, for some tumor typologies, metastasis occurs preferentially in a specific tissue type. This is known as the "seed-and-soil" hypothesis, meaning that cancer cell "seeds" travelling in the blood vessels prefer some specific tissues to metastasize as they are optimal "soil" for their growth. For example, skin tumour cells preferentially metastasize in the brain. As such, co-occurrence of tumor typologies in publications may mean that cells of a particular tumor typology that is the main subject of a publication, preferentially migrate to the tissue of the second tumor typology, co-occurring in the publication but not necessarily the main subject of the paper.

5 Evaluation

5.1 Experimental Setup

The aim of our evaluation was to show that the system presented above is well suited for the management of large volumes of Linked Data and can consequently support the extension of this data by novel RDF data extracted from other data streams. We thus compared TopFed with FedX [3] on 25 patients genomic results (clinical, methylation,SNP, exon-expression, gene-expression, miRNA, RNAseq2) extracted from 10 tumours. All of the data was distributed across 10 local SPARQL endpoints sharing a dedicated network.

We considered 10 queries of which 4 were star-shaped [1] and the remaining queries were path-shaped or hybrid (path+star) and contained between 3 and 7 triple patterns. We ran each of the queries 10 times and present the average

run-time for each of the queries. The query evaluation experiments were carried out on a 2.53GHz i5 processor with 4GB RAM. All of the data along with queries used for our evaluation can be found at TopFed home page.[9]

5.2 Results

As shown in Figure 3, we outperform FedX significantly on 90% of the queries. As an overall performance evaluation, the query run time of TopFed is about one third to that of FedX. While our best run-time (query 2, query 3) is more than 75 times smaller than that of FedX. We only have the same run-time for query 5. This is simply due to the number of sources selected by FedX being already optimal, thus efficient source selection based leading improvement was not possible. Due to smart source selection, we believe that our approach scales well on a large datasets query federation.

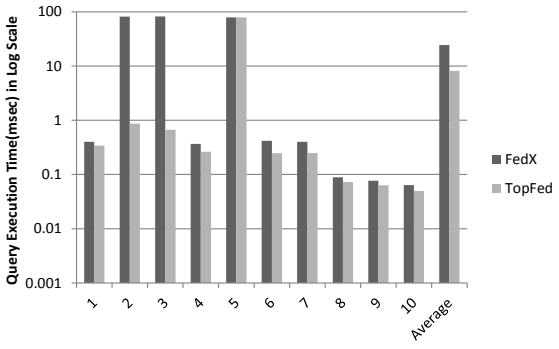

Fig. 3: Comparison of query runtimes

6 Conclusion and Future Work

In this paper, we presented a scalable solution for the continuous data integration coming from two large datasets. Further, we proposed a visual environment that aims to easily understand the data and support the serendipitous discovery of bio-medical hypotheses. We aim to integrate the complete Linked TCGA data (over 30 billions) with other digital libraries in future work. We will also explore new ways to further improve our data visualization platform.

References

1. Saleem, M., Ngonga Ngomo, A.C., Parreira, J.X., Deus, H.F., Hauswirth, M.: Daw: Duplicate-aware federated query processing over the web of data. In: International Semantic Web Conference (ISWC 2013). pp. 561–576 (2013)
2. Saleem, M., Shanmukha, S., Ngonga Ngomo, A.C., Almeida, J.S., Decker, S., Deus, H.F.: Linked cancer genome atlas database. In: I-Semantics 2013 (2013)
3. Schwarte, A., Haase, P., Hose, K., Schenkel, R., Schmidt, M.: Fedx: Optimization techniques for federated query processing on linked data. In: The Semantic Web, ISWC 2011, Lecture Notes in Computer Science, vol. 7031, pp. 601–616 (2011)

[9] `https://code.google.com/p/topfed/`

7 Appendix

In the following, we show how we fulfill each of the requirements of the Big Data Track.

7.1 Minimal Requirements

- Data Volume: Our implementation combines two very large datasets. The current version of Linked TCGA encompasses 7.6 billion triples with most of the quality control parameters stripped out. PubMed contains more than 23 million publications, each of which can be parsed into an arbitrary number of triples.
- Data Variety: We deal with both structured and unstructured data from diverse sources and in different formats. The Linked TCGA data was extracted from CSV files that were preprocessed, cleaned and transformed to RDF. Each of these files, however, though originating from the same source, are of very different nature as they measure many biological parameters - clinical diagnostic and outcome information, but also genetic and molecular information. We process the metadata associated with PubMed publications (especially the MESH annotations, title and authors) and transform them into RDF. Unstructured data (i.e., the publication abstracts) is processed to extract mentions of gene names and cancers.
- Data Velocity: the TCGA data doubles in size every two months. Similarly, the rate of new paper publication is in the order of 10000 per month[10]. Our framework can easily integrate novel data streams or data sources thanks to the underlying federated query engine.

7.2 Additional Desirable Features

- Usability: We devised an integrated visualization platform for the billions of triples underlying our application. This visualization allows bio-medical experts 1) to gain an overview of publications on bio-medical topics of interest, especially related to TCGA resources and 2) to formulate hypotheses w.r.t. to the interaction between cancers types, genes, drugs, etc.
- Value: By integrating Linked TCGA and related publications, discovered by our framework, annotated in PubMed, we create added value as we allow bio-medical experts not only to explore billions of triples but also to get a concise overview of the relations between the resources in the data set and the publications on these resources.
- Functionality: Our tool covers functionality centered around ingesting data streams from PubMed, extracting and connecting data from these streams to Linked TCGA and making the integrated data browse-able through an integrative visualization paradigm. We can serve novel data within a time scope suggested by the end user.

[10] Data from `http://www.nihms.nih.gov/stats/`